JOURNAL OF MULTIVARIATE ANALYSIS **59**, 187–205 (1996)
ARTICLE NO. 0060

Multivariate Locally Weighted Polynomial Fitting and Partial Derivative Estimation

Zhan-Qian Lu

Geophysical Statistics Project, National Center for Atmospheric Research, Boulder, Colorado 80307

Nonparametric regression estimator based on locally weighted least squares fitting has been studied by Fan and Ruppert and Wand. The latter paper also studies, in the univariate case, nonparametric derivative estimators given by a locally weighted polynomial fitting. Compared with traditional kernel estimators, these

locally weighted polynomial fitting to the estimation of partial derivatives in a multivariate regression context. Specifically, for both the regression and partial derivative estimators we prove joint asymptotic normality and derive explicit asymptotic expansions for their conditional bias and conditional convariance matrix (given observations of predictor variables) in each of the two important cases of local linear fit and local quadratic fit.

1. INTRODUCTION

Nonparametric regression estimation from a locally weighted least squares fit has been studied by Stone (1977, 1980), Cleveland (1979), Cleveland and Devlin (1988), Fan (1993), and Ruppert and Wand (1994). The last paper also studied, in the univariate case, nonparametric derivative estimators given by a locally weighted polynomial fitting. In this paper, we develop current work on locally weighted regression and we generalize locally weighted polynomial fitting the estimation of partial derivatives in a multivariate regression context.

We consider the following setup: Given i.i.d. random vectors, $\{(X_i, Y_i), i=1, ..., n\}$, where $Y_i \in \mathscr{R}^1$, $X_i \in \mathscr{R}^p$ and the latter has density function f. The statistical issues are estimation of the regression function

Received February 15, 1995; revised July 1996.

AMS 1991 subject classification: 62G07

Key words and phrases: locally weighted regression, joint asymptotic normality, asymptotic bias, asymptotic variance, kernel estimator, nonparametric regression.

0047-259X/96 $18.00

$m(\mathbf{x}) = E(Y|X=\mathbf{x})$ and its partial derivatives at any given point $\mathbf{x} \in \mathscr{R}^p$ in a domain of interest. Alternatively, if $EY^2 < \infty$ we can write

$$Y_i = m(X_i) + v^{1/2}(X_i)\,\varepsilon_i, \qquad i = 1, ..., n, \tag{1.1}$$

where ε_i's are i.i.d. scalar random variables with $E(\varepsilon_i | X_i) = 0$, $\mathrm{Var}(\varepsilon_i | X_i) = 1$, and v is called the variance function. This setup is called the *random design* model, as opposed to the *fixed design* model, given by

$$Y_i = m(\mathbf{x}_i) + v^{1/2}(\mathbf{x}_i)\,\varepsilon_i, \qquad i = 1, ..., n,$$

where the $\mathbf{x}_i$'s are predetermined quantities and nonrandom. In this paper, all results are given for the random design model, although similar results hold for the fixed design model as well.

The two kernel regression estimators, namely the *Nadaraya–Watson* (*N–W*) *estimator* (Nadaraya, 1964; Watson, 1964) and the *Gasser–Müller* (*G–M*) *estimator* (Gasser and Müller, 1979), have been studied extensively in the literature. Substantial recent attention focuses on a larger class of kernel estimators given by the locally weighted polynomial fitting. (The N–W estimator corresponds to the local constant fit). Chu and Marron (1991) found it difficult to compare the N–W and G–M estimators in a random design model. Subsequently Fan (1993) studied the nonparametric regression estimator based on the local linear fit and showed that the local linear regression smoother has advantages over both the N–W and G–M estimators as well as a certain optimality property. Another important issue is the boundary effect in the N–W and G–M estimators, namely both have slower convergence rates near boundary points and require some boundary modification for global estimation. As shown in Fan and Gijbels (1992) the local linear regression estimator does not have this drawback.

Derivative estimation often arises in bandwidth selection problems (Härdle, 1990) and in modeling growth curves in longitudinal studies (Müller, 1988). Our motivation for studying partial derivative estimation arises from a practical problem in chaos. Specifically, estimating Lyapunov exponents is an important issue which involves estimating first-order partial derivatives of a multivariate autoregression function. McCaffrey *et al.* (1992) have proposed applying nonparametric regression to this problem and they employed the derivative estimators given by differentiating certain nonparametric regression estimators.

Traditionally another often employed approach to derivative estimation is by kernel estimators using higher-order kernels. However, adaptations of these approaches for partial derivative estimation in a random design model are not very satisfactory since the derived estimators often have complicated form and are not easy to analyze. For example, the asymptotic bias and variance associated with the differentiation approach are difficult

to work out. Furthermore, these approaches suffer from drawbacks similar to those of the N–W and G–M estimators as discussed earlier.

We will show, generalizing Ruppert and Wand (1994), that the partial derivative estimators given by the local quadratic fit have desirable properties. The following results will be proved in this paper. In each of the two important cases of the local linear fit and local quadratic fit, for both the regression and derivative estimators, explicit asymptotic expansions are derived for their conditional bias and conditional covariance matrix, given observations of predictor variables (in Theorem 1 and Theorem 3). In each case, the joint asymptotic normality is also proved respectively (in Theorem 2 and Theorem 4). The results on bias and variance calculations of regression estimators in Theorem 1 and Theorem 3 correspond to those of Ruppert and Wand (1994), but more explicit expressions are given here. The results on derivative estimators generalize Ruppert and Wand (1994) to the multivariate regression case.

In Lu (1995b; see also Lu, 1994), the multivariate local polynomial fitting has been generalized to the multivariate time series context, where the joint asymptotic normality is proved, and the method is applied to estimating the spectra of local Lyapunov exponents.

Higher order fit, such as local cubic fit, can also be studied similarly, but the number of parameters to be estimated at each fitting increases very rapidly and may not be practical except for very large data set. It should be pointed out that, due to the curse of dimensionality in nonparametric smoothing methods, in order for the local polynomial fitting to be consistent, the sample size required grows exponentially fast as the dimension p increases. So for large p and a moderate amount of data some dimension reduction principle should be employed.

This paper is organized follows. Notations are given in Section 2. The locally weighted polynomial fit is introduced in Section 3, where the local linear fit is studied. Section 4 studies the main case of the local quadratic fit. The proofs of theorems are given in Section 5.

2. NOTATION

Given a $p \times p$ matrix A, A^{T} denotes its transpose. For $A_1, ..., A_k$ $(k > 1)$ which are square matrices, we denote

$$\operatorname{diag}\{A_1, ..., A_k\} = \begin{pmatrix} A_1 & & \\ & \ddots & \\ & & A_k \end{pmatrix},$$

where the suppressed elements are zeros.

For a given $p \times q$ matrix $A = (a_1, a_2, ..., a_q)$, let vec $A = (a_1^T, a_2^T, ..., a_q^T)^T$. If $A = (a_{ij})$ is a symmetric matrix of order p, let vech A denote the column vector of dimension $p(p+1)/2$, formed by stacking the elements on and below the diagonal; that is, vech $A = (a_{11}, ..., a_{p1}, a_{22}, ..., a_{p2}, ..., a_{pp})^T$. Also $\text{vech}^T A = (\text{vech } A)^T$.

Let U denote an open neighborhood of $\mathbf{x} = (x_1, ..., x_p)^T$ in $\mathscr{R}^p$, and let $C^d(U)$ be the class of functions which have up to order d continuous partial derivatives in U. For any $g = g(x_1, ..., x_p) \in C^d(U)$ and a positive number k (less than d), the kth-order differential $D_g^k(\mathbf{x}, \mathbf{u})$ for any given point $\mathbf{u} = (u_1, ..., u_p) \in \mathscr{R}^p$ is defined by

$$D_g^k(\mathbf{x}, \mathbf{u}) = \sum_{i_1, ..., i_p} C_{i_1 \cdots i_p}^k \frac{\partial^k g(\mathbf{x})}{\partial x_1^{i_1} \partial x_2^{i_2} \cdots \partial x_p^{i_p}} u_1^{i_1} \cdots u_p^{i_p},$$

where the summations are over all distinct nonnegative integers $i_1, ..., i_p$ such that $i_1 + \cdots + i_p = k$, and $C_{i_1 \cdots i_p}^k = k!/(i_1! \cdots i_p!)$. We also denote $D_g(\mathbf{x}) = (\partial g(\mathbf{x})/\partial x_1, ..., \partial g(\mathbf{x})/\partial x_p)^T$ for $g \in C^1(U)$ and the Hessian matrix by $\mathscr{H}_g(\mathbf{x}) = (\partial^2 g(\mathbf{x})/\partial x_i \partial x_j)$ for $g \in C^2(U)$.

For a $p \times p$ matrix A, we denote its determinant by $|A|$, and a certain norm by $\|A\|$, for example, $\|A\| = (\sum_{i,j=1}^p A(i,j)^2)^{1/2}$. Given a random sequence $\{a_n\}$, we denote $a_n = o_p(\gamma_n)$ if $\gamma_n^{-1} a_n$ tends to zero in probability, we denote $a_n = O_p(\gamma_n)$ if $\gamma_n^{-1} a_n$ tends to zero in probability, we denote $a_n = O_p(\gamma_n)$ if $\gamma_n^{-1} a_n$ is bounded in probability. For a sequence of $p \times q$ random matrices $\{A_n\}$, write $A_n = o_p(\gamma_n)$, or $O_p(\gamma_n)$ if and only if each component $A_n(i,j) = o_p(\gamma_n)$, or $O_p(\gamma_n)$, $i = 1, ..., p$, $j = 1, ..., q$. Then $A_n = o_p(\gamma_n)(O_p(\gamma_n))$ if and only if $\|A_n\| = o_p(\gamma_n)(O_p(\gamma_n))$.

3. LOCAL LINEAR FIT

The local linear estimators of regression and partial derivatives at any given $\mathbf{x}$ are given by the locally weighted least squares fit of a linear function, i.e., derived by minimizing the weighted sum of squares

$$\sum_{i=1}^n \{ Y_i - a - b^T(X_i - \mathbf{x}) \}^2 |H|^{-1} K(H^{-1}(X_i - \mathbf{x})), \tag{3.1}$$

where $K(\cdot)$ is the weighting function, H is the bandwidth matrix, and a and b are parameters.

We denote $Y = (Y_1, ..., Y_n)^T$,

$$W = \text{diag}\{ |H|^{-1} K(H^{-1}(X_1 - \mathbf{x})), ..., |H|^{-1} K(H^{-1}(X_n - \mathbf{x})) \},$$

and use $\mathbf{X}$ to denote the $n \times (p+1)$ design matrix

$$\mathbf{X} = \begin{pmatrix} 1 & (X_1 - \mathbf{x})^{\mathrm{T}} \\ \vdots & \vdots \\ 1 & (X_n - \mathbf{x})^{\mathrm{T}} \end{pmatrix}.$$

If there are at least $(p+1)$ points with positive weights, $(\mathbf{X}^{\mathrm{T}} W \mathbf{X})$ is invertible with probability one, and there is a unique solution to the minimization (3.1) given in matrix form by

$$\hat{\beta}_L = (\mathbf{X}^{\mathrm{T}} W \mathbf{X})^{-1} \mathbf{X}^{\mathrm{T}} W Y, \tag{3.2}$$

which is an estimator of $\beta_L(\mathbf{x}) = (m(\mathbf{x}), D_m^{\mathrm{T}}(\mathbf{x}))^{\mathrm{T}}$.

Use of a bandwidth matrix H in a multivariate smoothing context is sometimes advantageous as discussed by Ruppert and Wand (1994) (whose H corresponds to the H^2 here) and is also adopted in this paper following the suggestion of a referee.

We assume that:

(A1) The bandwidth matrix H is symmetric and strictly positive definite.

One choice is $H = hI$, where h is a scalar bandwidth and I is the identity matrix of dimension p. This choice implies that all predictor variables are scaled equally. However, since this special bandwidth matrix is simple and only one smoothing parameter needs to be specified, it remains the predominant choice in practice. If the predictor variables do not have the same scale, some transformation before smoothing, such as normalization by the respective standard deviations, may be advisable.

The weighting or kernel function K is generally a nonnegative integrable function.

For simplicity, we make the following assumptitions on the kernel function:

(A2) The kernel K is a spherically symmetric density function i.e., there exists a univariate function k such that $K(\mathbf{x}) = k(\|\mathbf{x}\|)$ for all $\mathbf{x} \in \mathscr{R}^p$. Furthermore, we will assume that the kernel K has eight-order marginal moment, i.e.,

$$\int u_1^8 K(u_1, ..., u_p)\, du_1 \cdots du_p < \infty.$$

Consequently, the odd-ordered moments of K and K^2, when they exist, are zero; i.e., for $l = 1, 2$

$$\int u_1^{i_1} u_2^{i_2} \cdots u_p^{i_p} K^l(\mathbf{u})\, d\mathbf{u} = 0 \qquad \text{if} \quad \sum_{j=1}^{p} i_j \quad \text{is odd.}$$

Also let $\mu_\ell = \int u_1^\ell K(\mathbf{u})\, d\mathbf{u}$, $J_\ell = \int u_p^{i_p} K^\ell(\mathbf{u})\, d\mathbf{u}$ for any nonnegative integers ℓ. Intuitively, by using a spherical symmetric kernel the weight is a function of Mahalanobis distance between data and the point of interest, and the bandwidth matrix H controls both the shape and scale of smoothing.

The following smoothness assumptions are made on the regression and design density:

(A3) For the given point $\mathbf{x} = (x_1, ..., x_p)^{\mathrm{T}}$ with $f(\mathbf{x}) > 0$, $v(\mathbf{x}) > 0$, there is an open neighborhood U of $\mathbf{x}$ such that $m \in C^3(U)$, $f \in C^1(U)$, $v \in C^0(U)$.

Since unconditional moments of the estimators considered here may not exist in general, following the tradition of Fan (1993) and Rupert and Wand (1994), we will work with their conditional moments (given observations of predictor variables X_i's). Their large-sample behavior as $\|H\| \to 0$, $n|H| \to \infty$ is studied in detail. We also establish joint asymptotic normality of the estimators.

The following theorem gives the asymptotic expansions for the conditional bias and conditional convariance matrix of the local linear estimators. Recall that $\mathscr{H}_m(\mathbf{x})$ denotes the Hessian matrix of m at $\mathbf{x}$, i.e., $(\partial^2 m(\mathbf{x})/\partial x_i \partial x_j)$.

THEOREM 1. *Under model* (1.1) *and assumptions* (A1)–(A3), *for* $\|H\| \to 0$, $n|H| \to \infty$ *as* $n \to \infty$, *the conditional bias of the local linear regression and derivative estimators given by* (3.2) *have the asymtotic expansions*

$$E\left\{\begin{pmatrix}\widehat{m(\mathbf{x})}\\ \widehat{D_m(\mathbf{x})}\end{pmatrix} - \begin{pmatrix} m(\mathbf{x})\\ D_m(\mathbf{x})\end{pmatrix}\,\middle|\, X_1, X_2, ..., X_n\right\}$$
$$B_L(\mathbf{x}, H) + \mathrm{diag}\{1, H^{-1}\}\ \|H\|^2\,[\,o_p(\|H\|) + O_p(\{n|H|\}^{-1/2})], \quad (3.3)$$

where

$$B_L(\mathbf{x}, H) = \mathrm{diag}\{1, H^{-1}\}\begin{pmatrix}\dfrac{1}{2}\mu_2\,\mathrm{Tr}(\mathscr{H}_m(\mathbf{x})\,H^2)\\ \dfrac{1}{3!\,\mu_2}\,b(m, H) + \dfrac{1}{2\mu_2 f(\mathbf{x})}\,b_1(m, H)\end{pmatrix}, \quad (3.4)$$

where

$$b(m, H) = \int \mathbf{u} D_m^3(\mathbf{x}, H\mathbf{u})\, K(\mathbf{u})\, d\mathbf{u}, \quad (3.5)$$

$$b_1(m, H) = \int \mathbf{u}[\mathbf{u}^{\mathrm{T}} H \mathscr{H}_m(\mathbf{x})\, H\mathbf{u}][D_f^{\mathrm{T}}(\mathbf{x})\, H\mathbf{u}]\, K(\mathbf{u})\, d\mathbf{u}$$
$$- \mu_2^2 H D_f(\mathbf{x})\,\mathrm{Tr}[\mathscr{H}_m(\mathbf{x})\, H^2].$$

The conditional variance-covariance matrix has the asymptotic expansion

$$\operatorname{Cov}\left\{\begin{pmatrix}\widehat{m(\mathbf{x})}\\ \widehat{D_m(\mathbf{x})}\end{pmatrix}\middle| X_1, X_2, \ldots, X_n\right\} = \frac{v(\mathbf{x})}{nf(\mathbf{x})\,|H|}\operatorname{diag}\{1, H^{-1}\}\left[\begin{pmatrix}J_0 & 0\\ 0 & (J_2/\mu_2^2)\end{pmatrix} + o_p(1)\right]\operatorname{diag}\{1, H^{-1}\}. \tag{3.6}$$

The regression result of Theorem 1 duplicates exactly Theorem 2.1 of Ruppert and Wand (1994) (hereafter R&W) and the derivative result is a new contribution of this paper and serves as a comparison to the local quadratic fit in the next section.

The joint asymptotic normality of the local linear estimators follows under the additional assumptition that

(A4) there eixsts a $\delta > 0$ such that $E\,|Y|^{2+\delta} < \infty$.

THEOREM 2. *Under conditions of Theorem* 1 *and* A4, *we have that*

$$(n\,|H|)^{1/2}\operatorname{diag}\{1, H\}\{\hat{\beta}_L - \beta(\mathbf{x}) - [B_L(\mathbf{x}, H) + \operatorname{diag}\{1, H^{-1}\}\, o(\|H\|^3)]\}$$

tends in distribution to $N(0, [v(\mathbf{x})/f(\mathbf{x})]\operatorname{diag}\{J_0, (J_2/u_2^2)\})$, *where* $B_L(\mathbf{x}, H)$ *is given by* (3.4).

Remark 1. For the results on the regression estimator alone to hold, weaker assumptions in (A3) such as $m \in C^2(U)$, and $f \in C^0(U)$ will suffice.

Remark 2. The convergence rate of the derivative estimator corresponding to $H = O(n^{-1/(p+6)})\,I$ is of order $n^{-2/(p+6)}$, which attains the optimal rate as established in Stone (1980). For $p = 1$, the conditional mean squared error of $\widehat{m'(x)}$ is approximated by

$$h^4\left\{\frac{\mu_4}{3!\,\mu_2}m^{(3)}(x) + \frac{\mu_4 - \mu_2^2}{2\mu_2}\frac{m^{(2)}(x)\,f'(x)}{f(x)}\right\}^2 + \frac{v(x)\,J_2}{\mu_2^2 f(x)\,nh^3}.$$

It can be checked that above expression corresponds to Theorem 4.2 of R&W (with $p = 1$, $r = 1$ in R&W's notation).

Remark 3. The bias of the local linear derivative estimator depends on the partial derivatives of the design density. This may be undesirable in some sense, e.g., in the minimax sense, analogous to criticism of the N-W estimator in Fan (1993). Furthermore, the local linear derivative estimators have boundary effect; i.e., at boundary points the asymptotic bias is of order $O(\|H\|)$.

4. LOCAL QUADRATIC FIT

Local quadratic fit is sometimes desirable for regression estimation as discussed in Cleveland and Devlin (1988). We will show that the local quadratic fit is suitable for first-order partial derivative estimation. The local quadratic estimators at a given point $\mathbf{x}$ are derived by minimizing the weighted sum of squares

$$\sum_{i=1}^{n} \{ Y_i - a - b^{\mathrm{T}}(X_i - \mathbf{x}) - (X_i - \mathbf{x})^{\mathrm{T}} L(X_i - \mathbf{x}) \}^2 \times |H|^{-1} K(H^{-1}(X_i - \mathbf{x})), \tag{4.1}$$

where a, b, L are parameters and L is restricted to be a lower triangular matrix for identifiability. Note that the total number of parameters in (4.1) is $q = \frac{1}{2}(p+2)(p+1)$. Denote Y, W as in Section 3,

$$\mathbf{X} = \begin{pmatrix} 1 & (X_1 - \mathbf{x})^{\mathrm{T}} & \mathrm{vech}^{\mathrm{T}}\{(X_1 - \mathbf{x})(X_1 - \mathbf{x})^{\mathrm{T}}\} \\ \vdots & \vdots & \vdots \\ 1 & (X_n - \mathbf{x})^{\mathrm{T}} & \mathrm{vech}^{\mathrm{T}}\{(X_n - \mathbf{x})(X_n - \mathbf{x})^{\mathrm{T}}\} \end{pmatrix}_{n \times q}.$$

If there are at least q points with positive weights in (4.1), then $\mathbf{X}^{\mathrm{T}} W \mathbf{X}$ is invertible with probability one, and there is a unique solution given in matrix form by

$$\hat{\beta} = (\mathbf{X}^{\mathrm{T}} W \mathbf{X})^{-1} \mathbf{X}^{\mathrm{T}} W Y, \tag{4.2}$$

which is an estimator of $\beta(\mathbf{x}) = (m(\mathbf{x}),\ D_m^{\mathrm{T}}(\mathbf{x}),\ \mathrm{vech}^{\mathrm{T}}\{(\mathbf{x})\})^{\mathrm{T}}$. Here $L(x) = (l_{ij})$ satisfies $l_{ij} = h_{ij}$ if $i > j$ and $h_{ii}/2$ if $i = j$, where $\mathscr{H}_m(\mathbf{x}) = (h_{ij})$ is the Hessian.

We will assume that:

(A5) The kernel K is as in (A2) and has 12th-order marginal moment, i.e.,

$$\int u_1^{12} K(u_1, ..., u_p)\, du_1 \cdots du_p < \infty.$$

(A6) For the given point $\mathbf{x}$ with $f(\mathbf{x}) > 0$, $v(\mathbf{x}) > 0$, there is an open neighborhood U of $\mathbf{x}$ such that $m \in C^4(U)$, $f \in C^1(U)$, $v \in C^0(U)$.

Additional notations are introduced. We define a square matrix $\mathscr{C}(H)$ of $p(p+1)/2$ such that $\mathrm{vech}\{H \mathbf{u}\mathbf{u}^{\mathrm{T}} H\} = \mathscr{C}(H)\, \mathrm{vech}\{\mathbf{u}\mathbf{u}^{\mathrm{T}}\}$ for any $\mathbf{u} \in \mathscr{R}^p$. Explicitly $\mathscr{C}(H) = L_1(H \otimes H)\, D_1$ where $\otimes$ denotes the Kronecker product and L_1 is the *elimination* matrix defined so that $L_1 \mathrm{vec}\, A = \mathrm{vec}\, A$ for any $p \times p$ matrix A, and D_1 is the *duplication* matrix defined so that $D_1 \mathrm{vech}\, A = \mathrm{vec}\, A$ for any symmetric matrix A. Further properties on these

special matrices are given in Magnus and Neudecker (1980). In particular, we have $\mathscr{C}(H)^i = L_1(H^i \otimes H^i)\, D_1$ for any $i = \cdots, -2, -1, 0, 1, 2, \ldots$. Letting $\|\cdot\|$ denote the matrix norm given by the largest singular value and using the fact that H is symmetric, we have that $\|\mathscr{C}(H)\| = \|H\|^2$. Hence $\|\mathscr{C}(H)\| = O(\|H\|^2)$ for any matrix norm $\|\cdot\|$ as $\|H\| \to 0$.

As in Section 3, we analyze the conditional bias and conditional covariance matrix of $\hat{\beta}$. The large sample asymptotic expansions are given in the next theorem.

THEOREM 3. *Under model* (1.1) *and assumptions* (A1), (A5), *and* (A6), *as* $\|H\| \to 0$ *and* $n|H| \to \infty$ *as* $n \to \infty$, *the conditional bias has the asymptotic expansion*:

$$E\left\{\begin{pmatrix} \widehat{m(\mathbf{x})} \\ \widehat{D_m(\mathbf{x})} \\ \operatorname{vech}\{\widehat{L(\mathbf{x})}\} \end{pmatrix} - \begin{pmatrix} m(\mathbf{x}) \\ D_m(\mathbf{x}) \\ \operatorname{vech}\{L(\mathbf{x})\} \end{pmatrix} \,\middle|\, X_1, X_2, \ldots, X_n\right\}$$
$$= B(\mathbf{x}, H) + \operatorname{diag}\{1, H^{-1}, \mathscr{C}(H)^{-1}\}\, \|H\|^3$$
$$\times [o_p(\|H\|) + O_p(\{n|H|\}^{-1/2})], \tag{4.3}$$

where

$$B(\mathbf{x}, H) = \operatorname{diag}\{1, H^{-1}, \mathscr{C}(H)^{-1}\} \begin{pmatrix} \frac{1}{4!}\,\theta(m, H) + \frac{1}{3!\, f(\mathbf{x})}\,\theta_1(m, H) \\ \frac{1}{3!\,\mu_2}\, b(m, H) \\ \frac{1}{4!}\,\gamma(m, H) + \frac{1}{3!\, f(\mathbf{x})}\,\gamma_1(m, H) \end{pmatrix}, \tag{4.4}$$

where $b(m, H)$ *is as in Theorem* 1, *and*

$$\theta(m, H) = d^{-1} \int D_m^4(\mathbf{x}, H\mathbf{u})\, K(\mathbf{u})\, d\mathbf{u} - c\, \operatorname{vech}\{I\} \int \operatorname{vech}\{\mathbf{u}\mathbf{u}^{\mathrm{T}}\}$$
$$\times D_m^4(\mathbf{x}, H\mathbf{u})\, K(\mathbf{u})\, d\mathbf{u},$$

$$\theta_1(m, H) = d^{-1} \int D_m^3(\mathbf{x}, H\mathbf{u})[D_f^{\mathrm{T}}(\mathbf{x})\, H\mathbf{u}]\, K(\mathbf{u})\, d\mathbf{u}$$
$$- c\, \operatorname{vech}\{I\} \int \operatorname{vech}\{\mathbf{u}\mathbf{u}^{\mathrm{T}}\}\, D_m^3(\mathbf{x}, H\mathbf{u})[D_f^{\mathrm{T}}(\mathbf{x})\, H\mathbf{u}]\, K(\mathbf{u})\, d\mathbf{u},$$

$$\gamma(m, H) = -c \operatorname{vech}\{I\} \int D_m^4(\mathbf{x}, H\mathbf{u})\, K(\mathbf{u})\, d\mathbf{u}$$

$$+ E^{-1} \int \operatorname{vech}\{\mathbf{uu}\}\, D_m^4(\mathbf{x}, H\mathbf{u})\, K(\mathbf{u})\, d\mathbf{u},$$

$$\gamma_1(m, H) = -c \operatorname{vech}\{I\} \int D_m^3(\mathbf{x}, H\mathbf{u})[D_f^{\mathrm{T}}(\mathbf{x})\, H\mathbf{u}]\, K(\mathbf{u})\, d\mathbf{u}$$

$$- Q^{\mathrm{T}}(\mathbf{x}, H) \int \mathbf{u} D_m^3(\mathbf{x}, H\mathbf{u})\, d\mathbf{u} + E^{-1} \int \operatorname{vech}\{\mathbf{uu}\}$$

$$\times D_m^3(\mathbf{x}, H\mathbf{u})[D_f^{\mathrm{T}}(\mathbf{x})\, H\mathbf{u}]\, K(\mathbf{u})\, d\mathbf{u},$$

and

$$d = (\mu_4 - \mu_2^2)/(\mu_4 + (p-1)\,\mu_2^2), \qquad c = \mu_2/(\mu_4 - \mu_2^2),$$

$$E = \operatorname{diag}\{\mu_4 - \mu_2^2, \underbrace{\mu_2^2, ..., \mu_2^2}_{p-1}, \mu_4 - \mu_2^2, \underbrace{\mu_2^2, ..., \mu_2^2}_{p-2}, ...,$$

$$\mu_4 - \mu_2^2, \mu_2^2, \mu_4 - \mu_2^2\}, \tag{4.5}$$

$$Q(\mathbf{x}, H) = -cHD_f(\mathbf{x}) \operatorname{vech}^{\mathrm{T}}\{I\}$$

$$+ \mu_2^{-1} \left\{ \int \mathbf{u} \operatorname{vech}^{\mathrm{T}}\{\mathbf{uu}^{\mathrm{T}}\}[D_f^{\mathrm{T}}(\mathbf{x})\, H\mathbf{u}]\, K(\mathbf{u})\, d\mathbf{u} \right\} E^{-1}.$$

The conditional variance–covariance matrix has asymptotic expansion,

$$\operatorname{Cov}\left\{ \begin{pmatrix} \widehat{m(\mathbf{x})} \\ \widehat{D_m(\mathbf{x})} \\ \operatorname{vech}\{\widehat{L(\mathbf{x})}\} \end{pmatrix} \middle| X_1, X_2, ..., X_n \right\}$$

$$= \frac{1}{n\,|H|} \operatorname{diag}\{1, H^{-1}, \mathscr{C}(H)^{-1}\} \{\Sigma(\mathbf{x}) + o_p(1)\}$$

$$\times \operatorname{diag}\{1, H^{-1}, \mathscr{C}^{\mathrm{T}}(H)^{-1}\}, \tag{4.6}$$

where

$$\Sigma(\mathbf{x}) = \frac{v(\mathbf{x})}{f(\mathbf{x})}$$

$$\begin{pmatrix} \rho & 0 & \phi \operatorname{vech}^{\mathrm{T}}\{I\} \\ 0 & J_2 \mu_2^{-2} I & 0 \\ \phi \operatorname{vech}\{I\} & 0 & \Lambda - \dfrac{\mu_2(J_2 - J_0\mu_2)}{(\mu_4 - \mu_2^2)^2} \operatorname{vech}\{I\} \operatorname{vech}^{\mathrm{T}}\{I\} \end{pmatrix}, \tag{4.7}$$

where

$$\rho = (\mu_4 - \mu_2^2)^{-2} \{J_0(\mu_4 + (p-1)\,\mu_2^2)^2 - 2pJ_2\mu_2(\mu_4 + (p-1)\,\mu_2^2) + p\mu_2^2(J_4 + (p-1)\,J_2^2)\},$$

$$\phi = (\mu_4 - \mu_2^2)^{-2} \{J_2\mu_4 + (2p-1)\,J_2\mu_2^2 - (p-1)\,J_2^2\mu_2 - J_4\mu_2 - J_0\mu_2\mu_4 - (p-1)\,J_0\mu_2^3\},$$

$$\Lambda = \operatorname{diag}\{\lambda_1, \underbrace{\lambda_2, ..., \lambda_2}_{p-1}, \lambda_1, \underbrace{\lambda_2, ..., \lambda_2}_{p-2}, ..., \lambda_1, \lambda_2, \lambda_1\},$$

where $\lambda_1 = (J_4 - J_2^2)(\mu_4 - \mu_2^2)^{-2}$, $\lambda_2 = J_2^2\mu_2^{-4}$.

Further, in regard to the matrix $Q(\mathbf{x}, H)$ defined in (4.5), it can be checked that the following result holds.

LEMMA 1. *Let* $(a_1, a_2, ..., a_p)^{\mathrm{T}} = HD_f(\mathbf{x})$, *the matrix* $Q(\mathbf{x}, H)$ *of* (4.5) *has the form*

$$\mu_2^{-1}\begin{pmatrix} a_1 & a_2 & \cdots & a_{p-1} & a_p & & & & & & & & \\ & a_1 & & & & a_2 & \cdots & a_{p-1} & a_p & & & & \\ & & \ddots & & & & \ddots & & & \cdots & & & \\ & & & a_1 & & & & a_2 & & & a_{p-1} & a_p & \\ & & & & a_1 & & & & a_2 & & & a_{p-1} & a_p \end{pmatrix},$$

where the suppressed elements are zeros.

Proof. Let $G(\mathbf{x}, H) = \int \mathbf{u}\,\operatorname{vech}\{\mathbf{u}\mathbf{u}^{\mathrm{T}}\}[D_f^{\mathrm{T}}(\mathbf{x})\,H\mathbf{u}]\,K(\mathbf{u})\,d\mathbf{u}$, which has the form:

$$\mu_2^2\begin{pmatrix} \frac{\mu_4}{\mu_2^2}a_1 & a_2 & \cdots & a_p & a_1 & 0 & \cdots & 0 & \cdots & a_1 & 0 & a_1 \\ a_2 & a_1 & \cdots & 0 & \frac{\mu_4}{\mu_2^2}a_2 & a_3 & \cdots & a_p & \cdots & a_2 & 0 & a_2 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & & \vdots & \vdots & \vdots \\ a_p & 0 & \cdots & a_1 & a_p & 0 & \cdots & a_2 & \cdots & a_p & a_{p-1} & \frac{\mu_4}{\mu_2^2}a_p \end{pmatrix}. \quad (4.8)$$

It can be checked that $Q(\mathbf{x}, H)$ has the given form. ∎

The regression part of Theorem 3 corresponds to the results in Section 3 of R&W. The derivative part of this theorem in the multivariate case is new and is the main contribution in this paper.

In order to see more clearly the connections between the results in this paper and the corresponding results in R&W, we give the explicit calculations of the matrix $A(H)$ (corresponding to N_x of R&W) which is used in our proofs. Define

$$A(H)=\int \begin{pmatrix} 1 \\ \mathbf{u} \\ \operatorname{vech}\{\mathbf{u}\mathbf{u}^{\mathrm{T}}\} \end{pmatrix} (1\ \mathbf{u}\ \operatorname{vech}\{\mathbf{u}\mathbf{u}^{\mathrm{T}}\})\, K(\mathbf{u})\, f(\mathbf{x}+H\mathbf{u})\, d\mathbf{u}.$$

We have the following lemma, whose proof is given in Section 5.

LEMMA 2. *If $f\in C^1(U)$ and $f(\mathbf{x})>0$, as $\|H\|\to 0$, we have*

$$A^{-1}(H)$$

$$=\frac{1}{f(\mathbf{x})}\begin{pmatrix} d^{-1} & 0 & -c\ \operatorname{vech}^{\mathrm{T}}\{I\} \\ 0 & \mu_2^{-1} & -f(\mathbf{x})^{-1}\, Q(\mathbf{x}, H) \\ -c\ \operatorname{vech}\{I\} & -f(\mathbf{x})^{-1}\, Q(\mathbf{x}, H)^{\mathrm{T}} & E^{-1} \end{pmatrix}+o(\|H\|),$$

where d, c, E, $Q(\mathbf{x}, H)$ are as in Theorem 3.

Explicit expressions for Theorem 3 and Theorem 4 are available in the important case that $H=hI$.

COROLLARY 1. *In the special case $H=hI$, we have that $\mathscr{C}(H)=h^2 I_2$ and $|H|=h^p$, where I_2 is the identity matrix of dimension $p(p+1)/2$, and the bias expressions in* (4.4) *have the explicit forms*

$$b(m, H)=h^3\begin{pmatrix} \mu_4\dfrac{\partial^3 m(\mathbf{x})}{\partial x_1^3}+3\mu_2^2\displaystyle\sum_{i=2}^{p}\frac{\partial^3 m(\mathbf{x})}{\partial x_i^2\,\partial x_1} \\ \mu_4\dfrac{\partial^3 m(\mathbf{x})}{\partial x_2^3}+3\mu_2^2\displaystyle\sum_{i\neq 2}\frac{\partial^3 m(\mathbf{x})}{\partial x_i^2\,\partial x_2} \\ \vdots \\ \mu_4\dfrac{\partial^3 m(\mathbf{x})}{\partial x_p^3}+3\mu_2^2\displaystyle\sum_{i=1}^{p-1}\frac{\partial^3 m(\mathbf{x})}{\partial x_i^2\,\partial x_p} \end{pmatrix} \tag{4.9}$$

and

$$\theta(m, H) = h^4 \left[\frac{\mu_4^2 - \mu_2\mu_6}{\mu_4 - \mu_2^2} \sum_{i=1}^{p} \frac{\partial^4 m(\mathbf{x})}{\partial x_i^4} - 6\mu_2^2 \sum_{1 \leqslant i < j \leqslant p} \frac{\partial^4 m(\mathbf{x})}{\partial x_i^2 \, \partial x_j^2} \right],$$

$$\theta_1(m, H) = h^4 \left[\frac{\mu_4^2 - \mu_2\mu_6}{\mu_4 - \mu_2^2} \sum_{i=1}^{p} \frac{\partial^3 m(\mathbf{x})}{\partial x_i^3} \frac{\partial f(\mathbf{x})}{\partial x_i} - 3\mu_2^2 \sum_{\substack{1 < i, j \leqslant p \\ i \neq j}} \frac{\partial^3 m(\mathbf{x})}{\partial x_i \, \partial x_j^2} \frac{\partial f(\mathbf{x})}{\partial x_i} \right];$$

and $\gamma(m, H)$ *and* $\gamma_1(m, H)$ *are vectors of dimension* $p(p+1)/2$ *with components*

$$\gamma(m, H) = (\gamma_{11}, ..., \gamma_{p1}, \gamma_{22}, ..., \gamma_{p2}, ..., \gamma_{pp})^{\mathrm{T}},$$

$$\gamma_{ii} = h^4 \left[\frac{\mu_6 - \mu_2\mu_4}{\mu_4 - \mu_2^2} \frac{\partial^4 m(\mathbf{x})}{\partial x_i^4} + 6 \frac{\mu_2\mu_4}{\mu_4 - \mu_2^2} \sum_{\substack{1 \leqslant k \leqslant p \\ k \neq i}} \frac{\partial^4 m(\mathbf{x})}{\partial x_i^2 \, \partial x_k^2}, \text{ for } 1 \leqslant i \leqslant p \right].$$

$$\gamma_{ij} = h^4 \left[4 \frac{\mu_4}{\mu_2} \left\{ \frac{\partial^4 m(\mathbf{x})}{\partial x_i^3 \, \partial x_j} + \frac{\partial^4 m(\mathbf{x})}{\partial x_i \, \partial x_j^3} \right\} + 12\mu_2 \sum_{\substack{1 \leqslant k \leqslant p \\ k \neq i, j}} \frac{\partial^4 m(\mathbf{x})}{\partial x_k^2 \, \partial x_i \, \partial x_j} \right],$$

$$\text{for} \quad 1 \leqslant j < i \leqslant p,$$

and $\gamma_1(m, H) = (\gamma_{11}(1), ..., \gamma_{p1}(1), \gamma_{22}(1), ..., \gamma_{p2}(1), ..., \gamma_{pp}(1))^{\mathrm{T}}$,

$$\gamma_{ii}(1) = h^4 \left[\frac{\mu_2\mu_6 - \mu_4^2}{\mu_2(\mu_4 - \mu_2^2)} \frac{\partial^3 m(\mathbf{x})}{\partial x_i^3} \frac{\partial f(\mathbf{x})}{\partial x_i} + 3\mu_2 \sum_{\substack{1 \leqslant k \leqslant p \\ k \neq i}} \frac{\partial^3 m(\mathbf{x})}{\partial x_i^2 \, \partial x_k} \right]$$

$$\text{for} \quad 1 \leqslant i \leqslant p.$$

$$\gamma_{ij}(1) = h^4 \left[6\mu_2 \sum_{\substack{1 \leqslant k \leqslant p \\ k \neq i, j}} \frac{\partial^3 m(\mathbf{x})}{\partial x_k \, \partial x_i \, \partial x_j} \frac{\partial f(\mathbf{x})}{\partial x_k} - 3\mu_2 \left\{ \frac{\partial^3 m(\mathbf{x})}{\partial x_i \, \partial x_j^2} \frac{\partial f(\mathbf{x})}{\partial x_j} + \frac{\partial^3 m(\mathbf{x})}{\partial x_i^2 \, \partial x_j} \frac{\partial f(\mathbf{x})}{\partial x_i} \right\} \right] \qquad \text{for} \quad 1 \leqslant j < i \leqslant p.$$

The joint asymptotic normality of the local quadratic estimators is given as follows.

THEOREM 4. *Under conditions of Theorem* 3 *and* (A4), *we have that*

$$(n|H)^{1/2} \operatorname{diag}\{1, H, \mathscr{C}(H)\} \times \{\hat{\beta} - \beta(\mathbf{x}) - [B(\mathbf{x}, H) + \operatorname{diag}\{1, H^{-1}, \mathscr{C}(H)^{-1}\} \, o(\|H\|^4)]\}$$

tends in distribution to $N(0, \Sigma(\mathbf{x}))$, *where* $B(\mathbf{x}, H)$ *and* $\Sigma(\mathbf{x})$ *are given as in* (4.4) *and* (4.7) *of Theorem* 3, *respectively.*

The following remarks are relevant:

Remark 4. For the results on the first-order partial derivatives to hold, the assumptions in (A6) can be weakened to $m \in C^3(U)$, $f \in C^0(U)$.

Remark 5. It is seen that the local quadratic derivative estimator eliminates the extra bias term $\{h^2/(2\mu_2 f(\mathbf{x}))\}\, b_1(m, K)$ of the local linear derivative estimator while retaining the asymptotic covariance matrix.

Remark 6. In the special case $H = hI$, the conditional mean squared distance of error (CMSDE) for the local quadratic gradient estimator $D_m(\mathbf{x})$ is given by

$$E\{\|\widehat{D_m(\mathbf{x})} - D_m(\mathbf{x})\|^2 \mid X_1, X_2, \ldots, X_n\} \approx \frac{h^4}{(3!\,\mu_2)^2}\|b(m)\|^2 + \frac{pJ_2 v(\mathbf{x})}{\mu_2^2 n h^{p+2} f(\mathbf{x})}, \tag{4.10}$$

where $b(m) = b(m, H)/h^3$. The locally optimal h which minimizes (4.10) is given by

$$h_{\text{opt}}(\mathbf{x}) = \left\{\frac{9p(p+2)\, J_2 v(\mathbf{x})}{f(\mathbf{x})\, \|b(m)\|^2}\right\}^{1/(p+6)} n^{-1/(p+6)}.$$

The minimum pointwise CMSDE is given by plugging in $h_{\text{opt}}(\mathbf{x})$,

$$\text{CMSDE}_{\text{opt}}(\mathbf{x}) = \frac{\{p(p+2)\, J_2 v(\mathbf{x})\}^{4/(p+6)}\, \|b(m)\|^{(2p+4)/(p+6)}}{3^{(2p+4)/(p+6)} 4\mu_2^2 f(\mathbf{x})^{4/(p+6)}} n^{-4/(p+6)}.$$

The asymptotic analysis provides some insights on the behavior of the estimator. The bias is quantified by the amount of smoothing and the third-order partial derivatives at $\mathbf{x}$ for each coordinate. Bias is increased when there is more third-order nonlinearity quantified by $b(m, K)$ and more smoothing. On the other hand, the conditional variance will be increased when there is less smoothing and sparser data near $\mathbf{x}$.

5. PROOFS

We only give the outlines of proofs of Theorem 3 and Theorem 4. The proof of Theorem 3 follows along the same lines as Fan (1993) and Ruppert and Wand (1994), and readers can refer to Lu (1995a) for a more detailed proof. Theorem 1 and Theorem 2 can be proved similarly.

5.1. *Outline of the Proof of Theorem* 3

The conditional bias and conditional covariance matrix are given respectively by

$$\begin{aligned} E(\hat{\beta} \mid X_1, ..., X_n) - \beta(\mathbf{x}) &= (\mathbf{X}^{\mathrm{T}} W \mathbf{X})^{-1} \mathbf{X}^{\mathrm{T}} W (M - \mathbf{X}\beta) \\ &= \operatorname{diag}\{1, H^{-1}, \mathscr{C}(H)^{-1}\}\, S_n^{-1} R_n, \\ \operatorname{Cov}(\hat{\beta} \mid X_1, ..., X_n) &= (\mathbf{X}^{\mathrm{T}} W \mathbf{X})^{-1} \mathbf{X}^{\mathrm{T}} W V W \mathbf{X} (\mathbf{X}^{\mathrm{T}} W \mathbf{X})^{-1} \\ &= \frac{1}{n\,|H|} \operatorname{diag}\{1, H^{-1}, \mathscr{C}(H)^{-1}\}\, S_n^{-1} C_n S_n^{-1} \\ &\quad \times \operatorname{diag}\{1, H^{-1}, \mathscr{C}^{\mathrm{T}}(H)^{-1}\} \end{aligned}$$

where $M = (m(X_1), ..., m(X_n))^{\mathrm{T}}$, $V = \operatorname{diag}\{v(X_1), ..., v(X_n)\}$,

$$S_n = \frac{1}{n} \sum_{i=1}^{n} \tilde{\mathbf{X}} \tilde{\mathbf{X}}^{\mathrm{T}} |H|^{-1} K(H^{-1}(X_i - \mathbf{x})),$$

$$C_n = \frac{1}{n} \sum_{i=1}^{n} \tilde{\mathbf{X}} \tilde{\mathbf{X}}^{\mathrm{T}} v(X_i) |H|^{-1} K^2(H^{-1}(X_i - \mathbf{x})),$$

$$R_n = \frac{1}{n} \sum_{i=1}^{n} \tilde{\mathbf{X}} \left[m(X_i) - m(\mathbf{x}) - D_m^{\mathrm{T}}(\mathbf{x})(X_i - \mathbf{x}) - \frac{1}{2} (X_i - \mathbf{x})^{\mathrm{T}} \mathscr{H}_m(\mathbf{x})(X_i - \mathbf{x}) \right] |H|^{-1} K(H^{-1}(X_i - \mathbf{x})),$$

where

$$\tilde{\mathbf{X}} = \begin{bmatrix} 1 \\ H^{-1}(X_i - \mathbf{x}) \\ \operatorname{vech}\{H^{-1}(X_i - \mathbf{x})(X_i - \mathbf{x})^{\mathrm{T}} H^{-1}\} \end{bmatrix}.$$

The proof of the bias part of the theorem consists of combining the following key steps with Lemma 2

$$\begin{aligned} S_n &= A(H) + O_p(\{n\,|H|\}^{-1/2}), \\ S_n^{-1} &= A^{-1}(H) + O_p(\{n\,|H|\}^{-1/2}), \\ R_n &= \{R(\mathbf{x}, H) + \|H\|^3 [o(\|H\|) + O_p(\{n\,|H|\}^{-1/2})]\}, \end{aligned} \tag{5.1}$$

where

$$R(\mathbf{x}, H) = \frac{1}{3!}\begin{pmatrix} \int \sum_{i=1}^{p} D_m^3(\mathbf{x}, H\mathbf{u})[D_f^{\mathrm{T}}(\mathbf{x})\, H\mathbf{u}]\, K(\mathbf{u})\, d\mathbf{u} \\ f(\mathbf{x}) \int \mathbf{u} \sum_{i=1}^{p} D_m^3(\mathbf{x}, H\mathbf{u})\, K(\mathbf{u})\, d\mathbf{u} \\ h \int \mathrm{vech}\{\mathbf{u}\mathbf{u}^{\mathrm{T}}\} \sum_{i=1}^{p} D_m^3(\mathbf{x}, H\mathbf{u})\, K(\mathbf{u})[D_f^{\mathrm{T}}(\mathbf{x})\, H\mathbf{u}]\, d\mathbf{u} \end{pmatrix}$$
$$+\frac{f(\mathbf{x})}{4!}\begin{pmatrix} \int D_m^4(\mathbf{x}, H\mathbf{u})\, K(\mathbf{u})\, d\mathbf{u} \\ 0 \\ \int \mathrm{vech}\{\mathbf{u}\mathbf{u}^{\mathrm{T}}\}\, D_m^4(\mathbf{x}, H\mathbf{u})\, K(\mathbf{u})\, d\mathbf{u} \end{pmatrix}.$$

The proof of covariance part of the theorem consists of combining (5.1.), Lemma 2 with

$$C_n = C(\mathbf{x}) + O(\|H\|) + O_p((n\,|H|)^{-1/2}),$$

where

$$C(\mathbf{x}) = v(\mathbf{x})\, f(\mathbf{x}) \begin{pmatrix} J_0 & 0 & J_2\, \mathrm{vech}^{\mathrm{T}}\{I\} \\ 0 & J_2 I_1 & 0 \\ J_2\, \mathrm{vech}\{I\} & 0 & E_J + J_2^2\, \mathrm{vech}\{I\}\, \mathrm{vech}^{\mathrm{T}}\{I\} \end{pmatrix}$$

and

$$E_J = \mathrm{diag}\{J_4 - J_2^2, \underbrace{J_2^2, ..., J_2^2}_{p-1}, J_4 - J_2^2, \underbrace{J_2^2, ..., J_2^2}_{p-2}, ..., J_4 - J_2^2, J_2^2, J_4 - J_2^2\}.$$

5.2. *Proof of Lemma* 2

Using the Taylor expansion

$$f(\mathbf{x} + H\mathbf{u}) = f(\mathbf{x}) + D_f^{\mathrm{T}}(\mathbf{x})\, H\mathbf{u} + o(\|H\|), \qquad \text{as} \qquad \|H\| \to 0,$$

we obtain that

$$A(H) = f(\mathbf{x}) \begin{pmatrix} 1 & 0 & \mu_2\, \mathrm{vech}^{\mathrm{T}}\{I\} \\ 0 & \mu_2 I & 0 \\ \mu_2\, \mathrm{vech}\{I\} & 0 & D \end{pmatrix}$$
$$+ \begin{pmatrix} 0 & \mu_2 H D_f^{\mathrm{T}}(\mathbf{x}) & 0 \\ \mu_2 H D_f(\mathbf{x}) & 0 & G(\mathbf{x}, H) \\ 0 & G(\mathbf{x}, H)^{\mathrm{T}} & 0 \end{pmatrix} + O(\|H\|),$$

where

$$D = E + \mu_2^2 \operatorname{vech}\{I\} \operatorname{vech}^{\mathrm{T}}\{I\},$$

and E as in Theorem 3 and $G(\mathbf{x}, H)$ as in (4.8).

Denote $A(H) = A_0 + B(H) + o(\|H\|)$, where A_0, $B(H)$ corresponds to the matrices in the expansion of $A(H)$ above. Note that

$$A(H)^{-1} = A_0^{-1} + A_0^{-1} B(H)\, A_0^{-1} + o(\|H\|),$$

where

$$A_0^{-1} = \frac{1}{f(\mathbf{x})} \begin{pmatrix} d^{-1} & 0 & -c \operatorname{vech}^{\mathrm{T}}\{I\} \\ 0 & \mu_2^{-1} I & 0 \\ -c \operatorname{vech}\{I\} & 0 & E^{-1} \end{pmatrix},$$

using matrix inverse formulae such as those in page 33 of Rao (1973).

Now we only need to check that the second term $A_0^{-1} B(H)\, A_0^{-1}$, which is $-f(\mathbf{x})^{-2}$ multiplied by

$$\begin{pmatrix} d^{-1} & 0 & -cV^{\mathrm{T}} \\ 0 & \mu_2^{-1} I & 0 \\ -cV & 0 & E^{-1} \end{pmatrix} \times \begin{pmatrix} 0 & \mu_2 (HD_f)^{\mathrm{T}} & 0 \\ \mu_2 HD_f & 0 & G \\ 0 & G^{\mathrm{T}} & 0 \end{pmatrix} \begin{pmatrix} d^{-1} & 0 & -cV^{\mathrm{T}} \\ 0 & \mu_2^{-1} I & 0 \\ -cV & 0 & E^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} 0 & d^{-1}(HD_f)^{\mathrm{T}} - c\mu_2^{-1} V^{\mathrm{T}} G^{\mathrm{T}} & 0 \\ d^{-1} HD_f - c\mu_2^{-1} GV & 0 & Q \\ 0 & Q^{\mathrm{T}} & 0 \end{pmatrix},$$

where $V = \operatorname{vech}\{I\}$ and $G = G(\mathbf{x}, H)$ and $Q = Q(\mathbf{x}, H)$.

Since

$$G(\mathbf{x}, H)\, V = (\mu_4 + (p-1)\, \mu_2^2)\, HD_f,$$

which can easily be checked from the explicit expression for $G(\mathbf{x}, H)$ in Lemma 1, thus

$$\frac{c}{\mu_2} G(\mathbf{x}, H)\, V = \frac{\mu_4 + (p-1)\, \mu_2^2}{\mu_4 - \mu_2^2} D_f = \frac{1}{d} D_f;$$

i.e., $d^{-1} HD_f - c\mu_2^{-1} G(\mathbf{x}, H)\, V = 0$, $d^{-1}(HD_f)^{\mathrm{T}} - c\mu_2^{-1} V^{\mathrm{T}} G(\mathbf{x}, H)^{\mathrm{T}} = 0$. The lemma is thus proved. ∎

5.3. *Outline of the Proof of Theorem* 4

We write

$$(n|H|)^{1/2}\,\mathrm{diag}\{1, H, \mathscr{C}(H)\} \times \{\hat{\beta} - \beta(\mathbf{x}) - \mathrm{diag}\{1, H^{-1}, \mathscr{C}(H)^{-1}\}\, S_n^{-1} R_n\} = S_n^{-1} Z_n, \quad (5.2)$$

where

$$Z_n = (n|H|)^{-1/2} \sum_{i=1}^{n} \tilde{\mathbf{X}} K(H^{-1}(X_i - \mathbf{x}))\, v^{1/2}(X_i)\, \varepsilon_i.$$

The joint asymptotic normality of Z_n can be established using the Cramer–Wold device and Theorem 1.9.3 of Serfling (1980) under the additional assumption (A4). The theorem is then proved by combining with previous results (5.1) in Subsection 5.1, and replacing the LHS of (5.2) by the bias given in Theorem 3 using Slusky's theorem (Theorem 1.5.4 of Serfling (1980)).

ACKNOWLEDGMENTS

This paper is based on part of the author's doctoral thesis finished in the Department of Statistics, University of North Carolina, under the supervision of Professor Richard L. Smith. Revision of this work was supported by the NCAR Geophysical Statistics Project, sponsored by the National Science Foundation under Grant DMS93-12686. This work benefitted from discussions with Professors Steve Marron and Jianqing Fan. Constructive suggestions from two anonymous referees have led to this condensed and improved version.

REFERENCES

[1] Chu, C. K., and Marron, J. S. (1991). Choosing a kernel estimator (with discussions). *Statist. Sci.* **6**, No. 4 404–436.

[2] Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74**, No. 368 829–836.

[3] Cleveland, W., and Devlin, S. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *J. Amer. Statist. Assoc.* **83**, No. 403 596–610.

[4] Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann of Statist.* **21**, No. 1 196–216.

[5] Fan, J., and Gijbels, I. (1992). Variable bandwith and local linear regression smoothers. *Ann. of Statist.* **20**, No. 4 2008–2036.

[6] Gasser, T., and Müller, H. G. (1979). Kernel estimation of regression functions. *Smoothing Techniques for Curve Estimation.* Lecture Notes in Math., Vol. 757, pp. 23–68. Springer-Verlag New York.

[7] Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge Univ. Press, Boston.

[8] Magnus, J. R., and Neudecker, H. (1980). The elimination matrix: Some lemmas and applications. *SIAM J. Alg. Disc. Methods* **1**, No. 4 422–449.
[9] Lu, Z. Q. (1994). *Estimating Lyapunov Exponents in Chaotic Time Series with Locally Weighted Regression*. Ph.D. dissertation. University of North Carolina, Chapel Hill.
[10] Lu, Z. (1995a). Multivariate locally weighted polynomial fit and partial derivative estimation. Unpublished.
[11] Lu, Z. Q. (1995b). Statistical estimation of local Lyapunov exponents: Toward characterizing predictability in nonlinear systems. Submitted.
[12] McCaffrey, D., Nychka, D., Ellner, S., and Gallant, A. R. (1992). Estimating Lyapunov exponents with nonparametric regression. *J. Amer. Statist. Soc.* **87**, No. 419 682–695.
[13] Müller, H.-G. (1988). *Nonparametric Regression Analysis of Longitudinal Data*. Springer-Verlag, Berlin.
[14] Nadaraya, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9** 141–142.
[15] Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York.
[16] Ruppert, D., and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22**, No. 3 1346–1370.
[17] Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.
[18] Stone, C. J. (1977). Consistent nonparametric regression (with discussions). *Ann. of Statist.* **5**, No. 4 595–645.
[19] Stone, C. J. (1980). Optimal rate of convergence for nonparametric estimators. *Ann. of Statist.* **8**, No. 6 1348–1360.
[20] Watson, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A* **26** 359–372.